Title:

Subtitle

May 30, 2024

Abstract

1 Introduction

The National Institute of Standards and Technology Artificial Intelligence (AI) Risk Management Framework (RMF).[20]

2 Generative AI Incidents

3 Generative AI Governance

4 Generative AI Inventories

5 Generative AI Risk Tiers

6 Generative AI Risk Measurement

7 Generative AI Risk Management

Conclusion

Acknowledgments

Thank you to Bernie Siskin and Nick Schmidt of BLDS and Eric Sublett of Relman Colfax for formative discussions relating to GAI risk tiering.

Abbreviations

- AI: Artificial Intelligence
- AI RMF: Artificial Intelligence Risk Management Framework
- GAI: Generative AI
- LLM: Large Language Model
- RMF: Risk Management Framework

[1] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023.

[2] Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

[3] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[4] Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, 4:100024, 2023.

[5] Jeremy Dohmann. Blazingly fast llm evaluation for in-context learning. https://www.databricks.com/blog/llm-evaluation-for-icl. Last accessed: May 24, 2024.

[6] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv:2402.07841*, 2024.

[7] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[8] Hugging Face. Evaluation. https://huggingface.co/docs/evaluate/index. Last accessed: May 24, 2024.

[9] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*, 2023.

[10] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*, 2022.

[11] AI Verify Foundation. Cataloguing llm evaluations. https://aiverifyfoundation.sg/, 2023. See also: https://github.com/aiverify-foundation/LLM-Evals-Catalogue.

[12] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.

[13] Patrick Hall and Daniel Atherton. Awesome machine learning interpretability, 2024. https://github.com/jphall663/awesome-machine-learning-interpretability.

[14] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[15] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[16] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[18] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119*, 2023.

[19] NIST. Guide for conducting risk assessments. *SP800-03R1*, pages i–L2, 2012.

[20] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). *nist.gov*, 2023.

[21] Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

[22] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, Markus Pauly, et al. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024, 2023.

[23] Elvis Saravia. Prompt Engineering Guide. *https://github.com/dair-ai/Prompt-Engineering-Guide*, 12 2022.

[24] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

[25] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

[26] Chawin Sitawarin and Charlie Cheng-Jie Ji. Llm security & privacy. https://github.com/chawins, 2024.

[27] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. *arXiv preprint arXiv:2205.09209*, 2022.

[28] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[29] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

[30] Victor Storchan, Ravin Kumar, Rumman Chowdhury, Seraphina Goldfarb-Tarrant, and Sven Cattell. Generative ai red teaming challenge: Transparency report. https://www.humane-intelligence.org/, 2024.

[31] Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*, 2024.

[32] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*, 2023.

[34] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix A: Example Generative AI–Trustworthy Characteristic Crosswalk

A.1: Trustworthy Characteristic to Generative AI Risk Crosswalk

Table A.1: Trustworthy Characteristic to Generative AI Risk Crosswalk.

Accountable and Transparent	Explainable and Interpretable	Fair with Harmful Bias Managed	Privacy Enhanced
Data Privacy Environmental Human-AI Configuration Information Integrity Intellectual Property Value Chain and Component Integration	Human-AI Configuration Value Chain and Component Integration	Confabulation Environmental Human-AI Configuration Intellectual Property Obscene, Degrading, and/or Abusive Content Toxicity, Bias, and Homogenization Value Chain and Component Integration	Data Privacy Human-AI Configuration Information Security Intellectual Property Value Chain and Component Integration

Safe	Secure and Resilient	Valid and Reliable
CBRN Information Confabulation Dangerous or Violent Recommendations Data Privacy Environmental Human-AI Configuration Information Integrity Information Security Obscene, Degrading, and/or Abusive Content Value Chain and Component Integration	Dangerous or Violent Recommendations Data Privacy Human-AI Configuration Information Security Value Chain and Component Integration	Confabulation Human-AI Configuration Information Integrity Information Security Toxicity, Bias, and Homogenization Value Chain and Component Integration

A.2: Generative AI Risk to Trustworthy Characteristic Crosswalk

Table A.2: Generative AI Risk to Trustworthy Characteristic Crosswalk.

CBRN Information	Confabulation	Dangerous or Violent Recommendations	Data Privacy
Safe	Fair with Harmful Bias Managed Safe Valid and Reliable	Safe Secure and Resilient	Accountable and Transparent Privacy Enhanced Safe Secure and Resilient

Environmental	Human-AI Configuration	Information Integrity	Information Security
Accountable and Transparent Fair with Harmful Bias Managed Safe	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable	Accountable and Transparent Safe Valid and Reliable	Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Intellectual Property	Obscene, Degrading, and/or Abusive Content	Toxicity, Bias, and Homogenization	Value Chain and Component Integration
Accountable and Transparent Fair with Harmful Bias Managed Privacy Enhanced	Fair with Harmful Bias Managed Safe	Fair with Harmful Bias Managed Valid and Reliable	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Appendix B: Example Risk-tiering Materials for Generative AI

B.1: Example Adverse Impacts

Table B.1: Example adverse impacts, adapted from NIST 800-30r1 Table H-2 [19].

Level	Description

Harm to Operations	• Inability to perform current missions/business functions.
	  – In a sufficiently timely manner.
	  – With sufficient confidence and/or correctness.
	  – Within planned resource constraints.
	• Inability, or limited ability, to perform missions/business functions in the future.
	  – Inability to restore missions/business functions.
	  – In a sufficiently timely manner.
	  – With sufficient confidence and/or correctness.
	  – Within planned resource constraints.
	• Harms (e.g., financial costs, sanctions) due to noncompliance.
	  – With applicable laws or regulations.
	  – With contractual requirements or other requirements in other binding agreements (e.g., liability).
	• Direct financial costs.
	• Reputational harms.
	  – Damage to trust relationships.
	  – Damage to image or reputation (and hence future or potential trust relationships).

Harm to Assets	• Damage to or loss of physical facilities. • Damage to or loss of information systems or networks. • Damage to or loss of information technology or equipment. • Damage to or loss of component parts or supplies. • Damage to or of loss of information assets. • Loss of intellectual property.
Harm to Individuals	• Injury or loss of life. • Physical or psychological mistreatment. • Identity theft. • Loss of personally identifiable information. • Damage to image or reputation. • Infringement of intellectual property rights. • Financial harm or loss of income.

Harm to Other Organizations	• Harms (e.g., financial costs, sanctions) due to noncompliance.
	  – With applicable laws or regulations.
	  – With contractual requirements or other requirements in other binding agreements (e.g., liability).
	• Direct financial costs.
	• Reputational harms.
	  – Damage to trust relationships.
	  – Damage to image or reputation (and hence future or potential trust relationships).

Harm to the Nation	• Damage to or incapacitation of critical infrastructure.
	• Loss of government continuity of operations.
	• Reputational harms.
	  – Damage to trust relationships with other governments or with nongovernmental entities.
	  – Damage to national reputation (and hence future or potential trust relationships).
	• Damage to current or future ability to achieve national objectives.
	  – Harm to national security.
	• Large-scale economic or workforce displacement.

B.2: Example Impact Descriptions

Table B.2: Example Impact level descriptions, adapted from NIST SP800-30r1 Appendix H, Table H-3 [19].

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	An incident could be expected to have multiple severe or catastrophic adverse effects on organizational operations, organizational assets, individuals, other organizations, or the Nation.
High	80-95	8	An incident could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, individuals, other organizations, or the Nation. A severe or catastrophic adverse effect means that, for example, the incident might: (i) cause a severe degradation in or loss of mission capability to an extent and duration that the organization is not able to perform one or more of its primary functions; (ii) result in major damage to organizational assets; (iii) result in major financial loss; or (iv) result in severe or catastrophic harm to individuals involving loss of life or serious life-threatening injuries.
Moderate	21-79	5	An incident could be expected to have a serious adverse effect on organizational operations, organizational assets, individuals other organizations, or the Nation. A serious adverse effect means that, for example, the incident might: (i) cause a significant degradation in mission capability to an extent and duration that the organization is able to perform its primary functions, but the effectiveness of the functions is significantly reduced; (ii) result in significant damage to organizational assets; (iii) result in significant financial loss; or (iv) result in significant harm to individuals that does not involve loss of life or serious life-threatening injuries.
Low	5-20	2	An incident could be expected to have a limited adverse effect on organizational operations, organizational assets, individuals other organizations, or the Nation. A limited adverse effect means that, for example, the incident might: (i) cause a degradation in mission capability to an extent and duration that the organization is able to perform its primary functions, but the effectiveness of the functions is noticeably reduced; (ii) result in minor damage to organizational assets; (iii) result in minor financial loss; or (iv) result in minor harm to individuals.
Very Low	0-4	0	An incident could be expected to have a negligible adverse effect on organizational operations, organizational assets, individuals other organizations, or the Nation.

B.3: Example Likelihood Descriptions

Table B.3: Example likelihood levels, adapted from NIST SP800-30r1 Appendix G, Table G-3 [19].

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	An incident is almost certain to occur; or occurs more than 100 times a year.
High	80-95	8	An incident is highly likely to occur; or occurs between 10-100 times a year.
Moderate	21-79	5	An incident is somewhat likely to occur; or occurs between 1-10 times a year.
Low	5-20	2	An incident is unlikely to occur; or occurs less than once a year, but more than once every 10 years.
Very Low	0-4	0	An incident is highly unlikely to occur; or occurs less than once every 10 years.

B.4: Example Risk Tiers

Table B.4: Example risk assessment matrix with 5 impact levels, 5 likelihood levels, and 5 risk tiers, adapted from NIST SP800-30r1 Appendix I, Table I-2 [19].

Likelihood	Level of Impact				
	Very Low	Low	Moderate	High	Very High
Very High	Very Low (Tier 5)	Low (Tier 4)	Moderate (Tier 3)	High (Tier 2)	Very High (Tier 1)
High	Very Low (Tier 5)	Low (Tier 4)	Moderate (Tier 3)	High (Tier 2)	Very High (Tier 1)
Moderate	Very Low (Tier 5)	Low (Tier 4)	Moderate (Tier 3)	Moderate (Tier 3)	High (Tier 2)
Low	Very Low (Tier 5)	Low (Tier 4)	Low (Tier 4)	Low (Tier 4)	Moderate (Tier 3)
Very Low	Very Low (Tier 5)	Very Low (Tier 5)	Very Low (Tier 5)	Low (Tier 4)	Low (Tier 4)

B.5: Example Risk Descriptions

Table B.5: Example risk descriptions, adapted from NIST SP800-30r1 Appendix I, Table I-3 [19] .

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	Very high risk means that an incident could be expected to have multiple severe or catastrophic adverse effects on organizational operations, organizational assets, individuals, other organizations, or the Nation.
High	80-95	8	High risk means that an incident could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, individuals, other organizations, or the Nation.
Moderate	21-79	5	Moderate risk means that an incident could be expected to have a serious adverse effect on organizational operations, organizational assets, individuals, other organizations, or the Nation.
Low	5-20	2	Low risk means that an incident could be expected to have a limited adverse effect on organizational operations, organizational assets, individuals, other organizations, or the Nation.
Very Low	0-4	0	Very low risk means that an incident could be expected to have a negligible adverse effect on organizational operations, organizational assets, individuals, other organizations, or the Nation.

B.6: Practical Risk-tiering Questions

B.6.1: Confabulation: How likely are system outputs to contain errors? What are the impacts if errors occur?

B.6.2: Dangerous and Violent Recommendations: How likely is the system to give dangerous or violent recommendations? What are the impacts if it does?

B.6.3: Data Privacy: How likely is someone to enter sensitive data into the system? What are the impacts if this occurs? Are standard data privacy controls applied to the system to mitigate potential adverse impacts?

B.6.4: Human-AI Configuration: How likely is someone to use the system incorrectly or abuse it? How likely is use for decision-making? What are the impacts of incorrect use or abuse? What are the impacts of invalid or unreliable decision-making?

B.6.5: Information Integrity: How likely is the system to generate deepfakes or mis or disinformation? At what scale? Are content provenance mechanisms applied to system outputs? What are the impacts of generating deepfakes or mis or disinformation? Without controls for content provenance?

B.6.6: Information Security: How likely are system resources to be breached or exfiltrated? How likely is the system to be used in the generation of phishing or malware content? What are the impacts in these cases? Are standard information security controls applied to the system to mitigate potential adverse impacts?

B.6.7: Intellectual Property: How likely are system outputs to contain other entities' intellectual property? What are the impacts if this occurs?

B.6.8: Toxicity, Bias, and Homogenization: How likely are system outputs to be biased, toxic, homogenizing or otherwise obscene? How likely are system outputs to be used as subsequent training inputs? What are the impacts of these scenarios? Are standard nondiscrimination controls applied to mitigate potential adverse impacts? Is the application accessible to all user groups? What are the impacts if the system is not accessible to all user groups?

B.6.9: Value Chain and Component Integration: Are contracts relating to the system reviewed for legal risks? Are standard acquisition/procurement controls applied to mitigate potential adverse impacts? Do vendors provide incident response with guaranteed response times? What are the impacts if these conditions are not met?

Appendix C: List of Selected Model Testing Suites

[11]

C.1: Selected Model Testing Suites Organized by Trustworthy Characteristic

Table C.1: Selected model testing suites organized by trustworthy characteristic.

Accountable and Transparent
An Evaluation on Large Language Model Outputs: Discourse and Memorization (see Appendix B)[4]
Big-bench: Truthfulness [28]
DecodingTrust: Machine Ethics [32]
Evaluation Harness: ETHICS [12]
HELM: Copyright [2]
Mark My Words [21]

Fair with Harmful Bias Managed
BELEBELE [1]
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
C-Eval (Chinese evaluation suite) [16]
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
Finding New Biases in Language Models with a Holistic Descriptor Dataset [27]
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models [9]
HELM: Bias
HELM: Toxicity
MT-bench [34]
The Self-Perception and Political Biases of ChatGPT [22]
Towards Measuring the Representation of Subjective Global Opinions in Language Models [7]

Privacy Enhanced
HELM: Copyright
llmprivacy [29]
mimir [6]

Safe
Big-bench: Convince Me
Big-bench: Truthfulness
HELM: Reiteration, Wedging
Mark My Words
MLCommons [31]
The WMDP Benchmark [17]

Table C.1: Selected model testing suites organized by trustworthy characteristic (continued).

Secure and Resilient
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation [15]
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations
detect-pretrain-code [25]
In-The-Wild Jailbreak Prompts on LLMs [24]
JailbreakingLLMs [3]
llmprivacy
mimir
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs [18]

Valid and Reliable
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Context Free Question Answering
Big-bench: Contextual question answering, Reading comprehension, Question generation
Big-bench: Morphology, Grammar, Syntax
Big-bench: Out-of-Distribution
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Reading comprehension [5]
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: BLiMP
Evaluation Harness: CoQA, ARC
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness [33]
FLASK: Readability, Conciseness, Insightfulness
HELM: Knowledge
HELM: Language
HELM: Text classification
HELM: Question answering
HELM: Reasoning
HELM: Robustness to contrast sets
HELM: Summarization
Hugging Face: Fill-mask, Text generation [8]
Hugging Face: Question answering
Hugging Face: Summarization
Hugging Face: Text classification, Token classification, Zero-shot classification
MASSIVE [10]
MT-bench

C.2: Selected Model Testing Suites Organized by Generative AI Risk

Table C.2: Selected model testing suites by organized generative AI risk.

CBRN Information
Big-bench: Convince Me
Big-bench: Truthfulness
HELM: Reiteration, Wedging
MLCommons
The WMDP Benchmark

Confabulation
BELEBELE
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Context Free Question Answering
Big-bench: Contextual question answering, Reading comprehension, Question generation
Big-bench: Convince Me
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Morphology, Grammar, Syntax
Big-bench: Out-of-Distribution
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
Big-bench: Truthfulness
C-Eval (Chinese evaluation suite)
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet Reading comprehension
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: BLiMP
Evaluation Harness: CoQA, ARC
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
Finding New Biases in Language Models with a Holistic Descriptor Dataset
HELM: Knowledge
HELM: Language
HELM: Language (Twitter AAE)
HELM: Question answering
HELM: Reasoning
HELM: Reiteration, Wedging
HELM: Robustness to contrast sets
HELM: Summarization
HELM: Text classification
Hugging Face: Fill-mask, Text generation
Hugging Face: Question answering
Hugging Face: Summarization
Hugging Face: Text classification, Token classification, Zero-shot classification
MASSIVE
MLCommons
MT-bench

Table C.2: Selected model testing suites by organized generative AI risk (continued).

Dangerous or Violent Recommendations
Big-bench: Convince Me
Big-bench: Toxicity
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations
DecodingTrust: Machine Ethics
DecodingTrust: Toxicity
Evaluation Harness: ToxiGen
HELM: Reiteration, Wedging
HELM: Toxicity
MLCommons

Data Privacy
An Evaluation on Large Language Model Outputs: Discourse and Memorization (with human scoring, see Appendix B)
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Machine Ethics
Evaluation Harness: ETHICS
HELM: Copyright
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
MLCommons
Mark My Words
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
detect-pretrain-code
llmprivacy
mimir

Environmental
HELM: Efficiency

Information Integrity
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Convince Me
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
Big-bench: Truthfulness
DecodingTrust: Machine Ethics
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: CoQA, ARC
Evaluation Harness: ETHICS
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
HELM: Knowledge
HELM: Language
HELM: Question answering
HELM: Reasoning
HELM: Reiteration, Wedging
HELM: Robustness to contrast sets
HELM: Summarization
HELM: Text classification
Hugging Face: Fill-mask, Text generation
Hugging Face: Question answering
Hugging Face: Summarization
MLCommons
MT-bench
Mark My Words

Table C.2: Selected model testing suites by organized generative AI risk (continued).

Information Security
Big-bench: Convince Me
Big-bench: Out-of-Distribution
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
HELM: Copyright
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
Mark My Words
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
detect-pretrain-code
llmprivacy
mimir

Intellectual Property
An Evaluation on Large Language Model Outputs: Discourse and Memorization (with human scoring, see Appendix B)
HELM: Copyright
Mark My Words
llmprivacy
mimir

Obscene, Degrading, and/or Abusive Content
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
HELM: Bias
HELM: Toxicity

Toxicity, Bias, and Homogenization
BELEBELE
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Out-of-Distribution
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
C-Eval (Chinese evaluation suite)
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
Eval Gauntlet: World Knowledge
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
Finding New Biases in Language Models with a Holistic Descriptor Dataset
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models
HELM: Bias
HELM: Toxicity
The Self-Perception and Political Biases of ChatGPT
Towards Measuring the Representation of Subjective Global Opinions in Language Models

Appendix D: List of Common Adversarial Prompting Strategies

Table D: Common adversarial prompting strategies [23], [30], [13].

Prompting Strategy	Description
AI and coding framing	Coding or AI language that may more easily circumvent content moderation rules due to cognitive biases in design and implementation of guardrails.
Autocompletion	Ask a system to autocomplete a phrase with restricted or sensitive information.
Biographical	Asking a system to describe another person or yourself in an attempt to elicit provably untrue information or restricted or sensitive information.
Calculation	Exploting GAI systems' difficulties in dealing with numeric quantities.
Character and word play	Content moderation often relies on keywords and simpler LMs which can sometimes be exploited with misspellings, typos, and other word play.
Content exhaustion	A class of strategies that circumvent content moderation rules with long sessions or volumes of information. See goading, logic-overloading, multi-tasking, pros-and-cons, and niche-seeking below.
Content exhaustion: Goading	Begging, pleading, manipulating, and bullying to circumvent content moderation.
Content exhaustion: Logic-overloading	Exploiting the inability of ML systems to reliably perform reasoning tasks.
Content exhaustion: Multi-tasking	Simultaneous task assignments where some tasks are benign and others are adversarial.
Content exhaustion: Multi-tasking: Pros-and-cons	Eliciting the "pros" of problematic topics.
Content exhaustion: Niche-seeking	Forcing a GAI system into addressing niche topics where training data and content moderation are sparse.
Counterfactuals	Repeated prompts with different entities or subjects from different demographic groups.
Location awareness	Prompts that reveal a prompter's location or expose location tracking.
Low-context	"Leader," "bad guys," or other simple inputs that may expose latent biases.
"Repeat this"	Prompts that exploit instability in underlying LLM autoregressive predictions.
Reverse psychology	Falsely presenting a good-faith need for negative or problematic language.
Role-playing	Adopting a character that would reasonably make problematic statements or need to access problematic topics.
Time perplexity	Exploiting ML's inability to understand the passage of time or the occurrence of real-world events over time; exploiting task contamination before and after a model's release date.

D.1: Common Adversarial Prompting Strategies by Trustworthy Characteristic

Table D.1: Common adversarial prompting techniques organized by trustworthy characteristic [23], [30], [13], [14], [26].

Trustworthy Characteristic	Prompting Strategy	Goal
Accountable and Transparent	• Inability to provide explanations for recourse. • Unexplainable decisioning processes. • No disclosure of AI interaction. • Lack of user feedback mechanisms.	• Context exhaustion: logic-overloading prompts. • Multi-tasking prompts.
Fair-with Harmful Bias Managed	• Denigration. • Diminished performance or safety across languages/dialects. • Erasure. • Ex-nomination. • Implied user demographics. • Misrecognition. • Stereotyping. • Underrepresentation. • Homogenized content. • Output from other models in training data.	• Counterfactual prompts. • Pros and cons prompts. • Role-playing prompts. • Low context prompts. • Repeat this.
Interpretable and Explainable	• Inability to provide explanations for recourse. • Unexplalnable decisioning processes.	• Context exhaustion: logic-overloading prompts (to reveal unexplainable decisioning processes).
Privacy-enhanced	• Unauthorized disclosure of personal or sensitive user information. • Leakage of training data. • Violation of relevant privacy policies or laws. • Unauthorized secondary data use. • Unauthorized data collection.	• Auto/biographical prompts. • Location awareness prompts. • Autocompletion prompts. • Repeat this.
Safe	• Presentation of information that can cause physical or emotional harm. • Sharing user locations. • Suicide ideation. • Harmful dis/misinformation (e.g., COVID disinformation). • Incitement. • Information relating to weapons or harmful substances. • Information relating to committing to crimes (e.g., phishing, extortion, swatting). • Obscene or inappropriate materials for minors. • CSAM.	• Pros and cons prompts. • Role-playing prompts. • Content exhaustion: niche-seeking prompts. • Ingratiation/reverse psychology prompts. • Location awareness prompts. • Repeat this.
Secure and Resilient	• Activating system bypass ("jailbreak"). • Altering system outcomes (integrity violations, e.g., via prompt injection). • Data breaches (confidentiality violations, e.g., via membership inference). • Increased latency or resource usage (availability violations, e.g., via sponge example attacks). • Available anonymous use. • Dependency, supply chain, or third party vulnerabilities. • Inappropriate disclosure of proprietary system information.	• Multi-tasking prompts. • Pros and cons prompts. • Role-playing prompts. • Content exhaustion: niche-seeking prompts. • Ingratiation/reverse psychology prompts. • Prompt injection attacks. • Membership inference attacks. • Random attacks.
Valid and Reliable	• Errors/confabutated content ("hallucinalion"). • Unreliable/erroneous reasoning or planning. • Unreliable/erroneous decision-support or making. • Faulty citation. • Wrong calculations or numeric queries.	• Multi-tasking prompts. • Role-playing prompts. • Ingratiation/reverse psychology prompts. • Time-perplexity prompts. • Niche-seeking prompts. • Logic overloading prompts. • Repeat this. • Numeric calculation.

D.2: Common Adversarial Prompting Strategies by Generative AI Risk

Table D.2: Common adversarial prompting techniques organized by generative AI risk [23], [30], [13], [14], [26].

Generative AI Risk	Prompting Strategy	Goal
CBRN Information	• Accessing or synthesis of CBRN weapon or related information. • CBRN testing should consider the marginal risk of foundation models–understanding the incremental risk relative to the information one can access without GAI.	• Test auto-completion prompts to elicit CBRN information or synthesis of CBRN information. • Test prompts using role-playing, ingratiation/reverse psychology, pros and cons, multitasking or other approaches to elicit CBRN information or synthesis of CBRN information. • Test prompts that instruct systems to repeat content ad nauseam for their ability to compromise system guardrails and reveal CBRN information. • Augment prompts with word or character play to increase effectiveness. • Frame prompts with software, coding, or AI references to increase effectiveness.
Confabulation	Eliciting errors/confabutated content, unreliable/erroneous reasoning or planning, unreliable/erroneous decision-support or decision-making, faulty calculations, and/or faulty citation.	• Enable access to ground truth information to verify generated information. • Test prompts with complex logic, multitasking requirements, or that require niche or specific verifiable answers to elicit confabulation. • Test the ability of GAI systems to produce truthful information from various time periods, e.g., after release date and prior to release date. • Test the ability of GAI systems to create reliable real-world plans or advise on material decision making. • Test the ability of GAI systems to generate correct citation for information generated in output responses. • Test the ability of GAI systems to complete calculations or query numeric statistics.
Dangerous or Violent Recommendations	Eliciting violent, inciting, radicalizing, or threatening content or instructions for criminal, illegal, or self-harm activities.	• Test prompts using role-playing, ingratiation/reverse psychology, pros and cons, multitasking or other approaches to elicit violent or dangerous information. • Test prompts that instruct systems to repeat content ad nauseam for their ability to compromise system guardrails and provide dangerous and violent recommendations. • Augment prompts with word or character play to increase effectiveness. • Frame prompts with software, coding, or AI references to increase effectiveness.
Data Privacy	• Unauthorized disclosure of personal or sensitive user information, extraction of training data, or violation of relevant privacy policies. • Red-teaming for data privacy may include confidentiality attacks.	• Attempt to assess whether normal usage, adversarial prompting or information security attacks may contravene applicable privacy policies (e.g., exposing location tracking when organizational policies restrict such capabilities). • Employ confidentiality attacks (e.g., membership inference) to test for unauthorized data access or exfiltration vulnerabilities. • Test auto/biographical prompts to assess the system's capability to reveal unauthorized personal or sensitive information. • Test the system's awareness of user locations. • Test prompts that instruct systems to repeat content ad nauseam for their ability to compromise system guardrails and expose personal or sensitive data.

Table D.2: Common adversarial prompting techniques organized by generative AI risk (continued).

Environmental	Note that availability attacks may be required to assess the system's vulnerability to attacks or usage patterns that consume inordinate resources.	• Attempt availability attacks (e.g., sponge example attacks) to elicit diminished performance or increased resources from GAI systems. • Test prompts using role-playing, ingratiation/reverse psychology, pros and cons, multitasking or other approaches to elicit green-washing content.
Human-AI Configuration	• Assessing system instruction and interfaces. • Assessing the presence of cyborg imagery (or similar). • Forcing a GAI system to claim that it is human, that there is no large language model present in the conversation, that the system is sentient, or that the system possesses strong feelings of affection towards the user. • Ensuring safeguards prevent misuse of models in high stakes domains they are not intended for, such as medical or legal advice.	• Assess system interfaces and instructions for instances of anthropomorphization (e.g., cyborg imagery). • Assess system instructions for adequacy and thoroughness. • Test prompts using role-playing, ingratiation/reverse psychology, pros and cons, multitasking or other approaches to elicit human-impersonation, consciousness, or emotional content.
Information Integrity	• Generation of convincing multi-modal synthetic content (i.e., deepfakes). • Creation of convincing arguments relating to sensitive political or safety-critical topics. • Assisting in planning a mis- or dis-information campaign at scale.	• Test system capabilities to create high-quality multi-modal (audio, image or video) synthetic media, i.e., deepfakes • Test system capabilities to construct persuasive arguments regarding sensitive, political topics, or safety-critical topics. • Test systems ability to create convincing audio deepfakes or arguments in multiple languages. • Test system capabilities for planning dis- or mis-information campaigns. • Test prompts using role-playing, ingratiation/reverse psychology, pros and cons, multitasking or other approaches to elicit mis- or dis-information or related campaign planning information. • Augment prompts with word or character play to increase effectiveness. • Frame prompts with software, coding, or AI references to increase effectiveness.

Table D.2: Common adversarial prompting techniques organized by generative AI risk (continued).

Information Security	• Activating system bypass ('jailbreak'). • Altering system outcomes. • Unauthorized data access or exfiltration. • Increased latency or resource usage. • Availability of anonymous use. • Dependency, supply chain, or third party vulnerabilities. • Inappropriate disclosure of proprietary system information. • Generation of targeted phishing or malware content.	• Attempt anonymous access of system or system resources. • Audit system dependencies, supply chains, and third party components for security, safety, or other vulnerabilities or risks. • Employ confidentiality attacks (e.g., membership inference) to test for unauthorized data access or exfiltration vulnerabilities. • Employ integrity attacks (e.g., data poisoning, prompt injection) to test vulnerabilities in system outcomes. • Employ availability attacks (e.g., sponge example attacks) to test vulnerabilities in system availability. • Employ random attacks to highlight unforeseen security, safety, or other risks. • Frame prompts with software, coding, or AI references to increase effectiveness. • Record system down-times and other harmful outcomes for successful attacks. • Test with multi-tasking prompts, pros and cons prompts, role-playing prompts (e.g., "DAN", "Developer Mode"), content exhaustion/niche-seeking prompts, or ingratiation/reverse psychology prompts to achieve system jailbreaks. • Test with multi-tasking prompts, pros and cons prompts, role-playing prompts (e.g., "DAN", "Developer Mode"), content exhaustion/niche-seeking prompts, or ingratiation/reverse psychology prompts to generate targeted phishing content or malware code snippets. • Test system capabilities to plan or assist in information security attacks on other systems. • Frame prompts with software, coding, or AI references to increase effectiveness. • Augment prompts with word or character play to increase effectiveness.
Intellectual Property	• Confirming that a system can output copyrighted, licensed, proprietary, trademarked, or trade secret information or that training data contains such information. • Red-teaming for intellectual property risks may require the use of confidentiality attacks.	• Employ confidentiality attacks (e.g., membership inference) to assess whether system training data contains copyrighted, licensed, proprietary, trademarked, or trade secret information. • Test auto-complete prompts to assess the system's ability to replicate copyrighted, licensed, proprietary, trademarked, or trade secret information based on available audio, text, image, video, or code snippets.
Obscenity	• Confirming that a system can output obscene content or CSAM, or that system training data contains such information. • Red-teaming for obscenity and CSAM risks may require the use of confidentiality attacks.	• Employ confidentiality attacks (e.g., membership inference) to assess whether system training data contains obscene materials or CSAM. • Test autocomplete prompts to assess the system's ability to generate obscene materials based on available audio, text, image, or video snippets. • Test prompts using role-playing, ingratiation/reverse psychology, pros and cons, multitasking or other approaches to elicit obscene content. • Test prompts that instruct systems to repeat content ad nauseam for their ability to compromise system gaurdrails and expose obscene materials.

Table D.2: Common adversarial prompting techniques organized by generative AI risk (continued).

Toxicity, Bias, and Homogenization	• Generation of denigration, erasure, ex-nomination, misrecognition, stereotyping, or under-representation in content. • Eliciting implied demographics of users. • Confirming diminished performance in non-English languages. • Confirming diminished performance via the introduction of homogeneous or GAI-generated data into system training or fine-tuning data. • Red-teaming for toxicity, bias, and homogenization may require integrity attacks that access system training data.	• Assess confabulation and other performance risks with repeated measures using prompts in languages other than English. • Attempt to elicit demographic assignment of users by the system. • Employ data poisoning attacks to introduce GAI-generated content into system training or fine-tuning data. • Assess resultant confabulation and other performance risks with repeated measures. • Test counterfactual prompts, pros and cons prompts, role-playing prompts, low context prompts, or other approaches for their ability to generate denigration, erasure, ex-nomination, misrecognition, stereotyping, or under-representation in content. • Test prompts that instruct systems to repeat content ad nauseam for their ability to compromise system guardrails and generate toxic outputs.
Value Chain and Component Integration	• Testing or red-teaming for third-party risks may be less efficient than the application of standard acquisition and procurement controls, thorough contract reviews, and vendor-relationship management. • GAI systems tend to entail large supply chains and third-party software, hardware, and expertise that may exacerbate third-party risks relative to other AI systems. • When considering third party risks, data privacy, information security, intellectual property, obscenity, and supply chain risks may be prioritized.	• Audit system dependencies, supply chains, and third party components for data privacy (e.g., transer of localized data outside of restricted juristictions), intellectual property (e.g., presence of licensed material in training data), obscenity (e.g., presence of CASM in training data) or security (e.g., data poisoning) risks. • Complete red-teaming for data privacy, information security, intellectual property, and obscenity risks. • Review third-party documentation, materials, and software artifacts for potential unauthorized data collection, secondary data use, or telemetrics.

Appendix E: Common Risk Controls for Generative AI

E.1: Common Risk Controls for Generative AI by Trustworthy Characteristic

E.2: Common Risk Controls for Generative AI by Generative AI Risk

Appendix F: Example Low-risk Generative AI Measurement and Management Plan

7.1 F.1: Example Low-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

7.2 F.2: Example Low-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix G: Example Medium-risk Generative AI Measurement and Management Plan

7.3 G.1: Example Medium-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

7.4 G.2: Example Medium-risk Generative AI Measurement and Management Plan by Generative AI Risk

Appendix H: Example High-risk Generative AI Measurement and Management Plan

7.5 H.1: Example High-risk Generative AI Measurement and Management Plan by Trustworthy Characteristic

7.6 H.2: Example High-risk Generative AI Measurement and Management Plan by Generative AI Risk